



MÁSTER DE BIG DATA

TEMARIO

MODULO I – INICIO DEL MASTER

- Introducción al curso.
- Introducción al Big Data.
- Apache Spark y su integración con Scala.

MODULO II – PLATAFORMA DATABRICKS

- Databricks Community Edition.
- Creación de una cuenta en Databricks Community Edition.
- Entorno de trabajo en Databricks.
- Creación de un clúster en Databricks.
- Creación de un notebook en Databricks.
- Importación de datos a Databricks.

MODULO III – INTRODUCCIÓN A LOS RDD EN SPARK UTILIZANDO SCALA

- Spark Sesión.
- ¿Qué es un RDD?
- Diferentes formas de crear un RDD en Scala.
- Ejercicios prácticos.
- Resolución de los ejercicios.



MODULO IV – TRANSFORMACIONES EN SPARK RDD CON SCALA

- Transformaciones en un RDD.
- Función map.
- Función flatMap.
- Función filter.
- Función coalesce.
- Función repartition.
- Función reduceByKey.
- Ejercicios prácticos.
- Resolución de los ejercicios.

MODULO V – ACCIONES EN SPARK RDD CON SCALA

- Acciones en un RDD.
- Función reduce.
- Función count.
- Función collect.
- Funciones take, max y saveAsTextFile.
- Ejercicios prácticos.
- Resolución de los ejercicios.



MODULO VI – CARACTERISTICAS AVANZADAS DE RDD EN SCALA

- Almacenamiento en caché.
- Particionado.
- Mezcla de datos (shuffling).
- Broadcast variable.
- Accumulators.
- Ejercicios prácticos.
- Resolución de los ejercicios.

MODULO VII – INTRODUCCIÓN A SPARK SQL CON SCALA

- Creación de un DataFrame a partir de un RDD en Scala.
- Creación de un DataFrame a partir de fuentes de datos en Scala.
- Trabajo con columnas.
- Funciones select y selectExpr.
- Funciones filter y where.
- Funciones distinct y dropDuplicates.
- Funciones withColumn y withColumnRenamed.
- Funciones drop, sample y randomSplit.
- Trabajo con datos incorrectos o faltantes.
- Acciones sobre un DataFrame en Spark SQL.
- Escritura de DataFrames.



- Persistencia de DataFrames.
- Ejercicios prácticos.
- Resolución de los ejercicios.

MODULO VIII – SPARK SQL AVANZADO CON SCALA

- Agregaciones.
- Funciones count, countDistinct y approx_count_distinct.
- Funciones min y max.
- Funciones sum, sum_distinct y avg.
- Agregación con agrupación.
- Varias agregaciones por grupo.
- Agregación con pivote.
- Joins.
- Expresión join y tipos de join.
- Inner Join, Left Outer Join, Right Outer Join, Full Outer Join, Left Anti Join, Left Semi Join, Cross Join.
- Manejo de nombres de columna duplicados.
- Shuffle Hash Join y Broadcast Hash Join.
- Ejercicios prácticos.
- Resolución de los ejercicios.



<Asociación
Española de
Programadores
Informáticos/>

MODULO IX – FUNCIONES EN SPARK SQL CON SCALA

- Funciones de fecha y hora.
- Funciones para trabajo con strings.
- Funciones para trabajo con colecciones.
- Funciones when, coalesce y lit.
- Funciones definidas por el usuario (UDF).
- Funciones de ventana.
- Catalyst Optimizer.
- Ejercicios prácticos.
- Resolución de los ejercicios.

MODULO X – PROYECTO FIN DE MÁSTER

aeppi